НЕЛИНЕЙНАЯ КОРРЕЛЯЦИОННАЯ ЗАВИСИМОСТЬ

Различают следующие **виды нелинейной корреляции**:

1. параболическую,
2. гиперболическую,
3. экспоненциальную и другие.

Путь зависимость между признаками X и Y задана в виде корреляционной таблицы. Для определения типа нелинейной в системе координат на плоскости строят точки $M_i\,(X_i;Y_i)$.

Если точки в корреляционном поле располагаются вблизи некоторой **параболы**, то уравнение регрессии записывают в виде

$$\boxed{\hat{y}_x = a_0 + a_1 x + a_2 x^2}. \qquad (1)$$

Оценка a_0, a_1, a_2 для неизвестных параметров истинного уравнения регрессии находят по методу наименьших квадратов. Если опытные данные не сгруппированы в корреляционную, то оценки находят, решая систему нормальных уравнений

$$\boxed{\begin{cases} n a_0 + a_1[x] + a_2[x^2] = [y] \\ a_0[x] + a_1[x^2] + a_2[x^3] = [xy] \\ a_0[x^2] + a_1[x^3] + a_2[x^4] = [x^2 y] \end{cases}} \qquad (2)$$

Для сгруппированных значений признаков X и Y оценки a_0, a_1, a_2 находят, решая систему, нормальных уравнений

$$\boxed{\begin{cases} n a_0 + a_1[n_x x] + a_2[n_x x^2] = [n_x \bar{y}_x] \\ a_0[n_x x] + a_1[n_x x^2] + a_2[n_x x^3] = [n_x x \bar{y}_x] \\ a_0[n_x x^2] + a_1[n_x x^3] + a_2[n_x x^4] = [n_x x^2 \bar{y}_x] \end{cases}} \qquad (3)$$

Зависимость между X и Y может быть близкой к **гиперболической**.
В этом случае уравнение регрессии ищут в виде

$$\boxed{\hat{y}_x = a_0 + \frac{a_1}{x}} \tag{4}$$

Оценки a_0 и a_1 неизвестных параметров истинного уравнения регрессии находят по методу наименьших квадратов, решая систему нормальных уравнений

$$\boxed{\begin{cases} a_0 n + a_1\left[\frac{1}{x}\right] = [y] \\ a_0\left[\frac{1}{x}\right] + a_1\left[\frac{1}{x^2}\right] = \frac{y}{x} \end{cases}} \tag{5}$$

где $[1/x]$ — сумма величин, обратных значениям x,
$[1/x^2]$ сумма их квадратов,
$[y]$ — сумма величин y,
$[y/x]$ — сумма отношений значений y_i к x_i.

Если гиперболическая зависимость между признаками X и Y имеет вид

$$\hat{y}_x = \frac{1}{a_0 + a_1 x} \tag{6}$$

то оценки a_0 и a_1 находят, решая систему нормальных уравнений

$$\begin{cases} a_0 n + a_1 [x] = \left[\frac{1}{y}\right] \\ a_0 [x] + a_1 [x^2] = \left[\frac{x}{y}\right] \end{cases} \tag{7}$$

Если зависимость между признаками имеет **экспоненциальный характер**, то уравнение регрессии ищут в виде

$$\boxed{\hat{y}_x = a \cdot b^x} \tag{8}$$

Для определения оценок a и b входящих в уравнение регрессии решают систему нормальных уравнений

$$\begin{cases} \lg a[x] + \lg b\left[x^2\right] = [x \lg y] \\ n \lg a + \lg b[x] = [\lg y] \end{cases} \tag{9}$$

Если в корреляционном поле около построенных точек предполагается проведение разных по типу линий (параболы, гиперболы, экспоненты, логарифмики), то для выбора одной из них, характеризующей наилучшим образом зависимость между признаками X и Y, применяют либо метод конечных разностей, либо производят проверку необходимых условий.

ПРОВЕРКА НЕОБХОДИМЫХ УСЛОВИЙ

Проверку для выбора одной из предполагаемых нелинейных зависимостей проводят, пользуясь табл. 1.

Таблица 1

Необходимые условия	Вид формулы	Способ выравнивания (приведения к линейной зависимости)
$y\left(\frac{x_1+x_n}{2}\right)=\frac{y(x_1)+y(x_n)}{2}$	$y=ax+b$	
$y(\sqrt{x_1\cdot x_n}=\sqrt{y(x_1)\cdot y(x_n)}$	$y=a\cdot x^b$	$Y=AX+B$, $X=\lg x$, $Y=\lg x$, $A=\lg a$, $B=\lg b$
$y\left(\frac{x_1+x_n}{2}\right)=\sqrt{y(x_1)\cdot y(x_n)}$	$y=a\cdot b^x$	$Y=AX+B$, $X=x$, $Y=\lg y$, $A=\lg a$, $B=\lg b$
$y(\frac{2x_1x_n}{x_1+x_n})=\frac{y(x_1)+y(x_n)}{2}$	$y=a+\frac{b}{x}$	$Y=AX+B$, $X=x$, $Y=xy$
$y(\frac{x_1+x_n}{2})=\frac{2y(x_1)\cdot y(x_n)}{y(x_1)+y(x_n)}$	$y=\frac{1}{ax+b}$	$Y=AX+B$, $X=x$, $Y=\frac{1}{y}$
$y(\sqrt{x_1\cdot x_n})=\frac{y(x_1)+y(x_n)}{2}$	$y=a\lg x+b$	$Y=AX+B$, $X=\lg x$, $Y=y$

РАБОТА С ТАБЛИЦЕЙ. Если выполняется одно из условий первого столбца таблицы, то выбирают в качестве предполагаемой формулы соответствующую формулу, стоящую во втором столбце таблицы рассматриваемой строки. В третьем столбце указывается способ выравнивания, то есть приведения изучаемой зависимости к линейной. Если выравненные точки $(X_i;Y_i)$ хорошо ложатся на прямую, то указанную во втором столбец таблицы зависимость принимаем в качестве предполагаемой.

Если значения функции, вычисленные в первом столбце таблицы при выбранных значениях аргумента отсутствуют в таблице опытных данных, то их находят линейным интерполированием по формуле

$$\boxed{y(x) = y(x_1) + \frac{y(x_2) - y(x_1)}{x_2 - x_1}(x - x_1)} \qquad (10)$$

где x_1 и x_2 — два рядом стоящих значения признака X в таблице опытных данных, между которыми находится значение x, вычисленное по табл. 1 первого столбца.

Для всех предполагаемых формул по результатам первого столбца табл. 1 вычисляют отклонения Δ правой части от левой необходимого условия. Вычисленные отклонения Δ_i сравнивают и по наименьшему из них выбирают окончательно одну из формул.

МЕТОД КОНЕЧНЫХ РАЗНОСТЕЙ

Пусть в корреляционном поле могут быть проведены линии, описываемые уравнениями $y = ax^b$, $y = ae^{cx}(c = \ln b)$, $y = a + \frac{b}{x}$, $y = \frac{1}{ax+b}$, $y = a\lg x + b$. Все эти формулы содержат по два параметра и могут быть приведены к формуле $Y = AX + B$, пользуясь таблицей 1. Так как все зависимости, приведенные в таблице 1, сводятся к линейной, то для обоснования выбора формулы $Y = AX + B$ вычисляют отношения $\Delta Y / \Delta X$, ΔX и ΔY — конечные разности первого порядка. Аналитическим критерием выбора формулы по этому методу служит тот факт, что отношения $\Delta Y / \Delta X$ мало отличаются друг от друга для выбранной формулы.

Если предполагаемая формула имеет вид $y = ax^2 + bx + c$, то критерием выбора этой формулы являются незначительные отклонения по модулю конечных разностей второго порядка $\Delta^2 y$ от среднего значения этих разностей $\overline{\Delta^2 y}$. **Конечные разности** находят, пользуясь табл. 2.

Таблица 2

Таблица конечных разностей

x	y	Δy	$\Delta^2 y$
x_0	y_0		
		Δy_0	
x_1	y_1		$\Delta^2 y_0$
		Δy_1	
x_2	y_2		$\Delta^2 y_1$
		Δy_2	
x_3	y_3		$\Delta^2 y_2$
		Δy_3	
x_4	y_4		$\Delta^2 y_3$
		Δy_4	
x_5	y_5		…
		…	
…	…		

Определение силы криволинейной связи

Для определения тесноты связи между признаками X и Y при нелинейной корреляции используют **корреляционные отношения и индекс корреляции.**

Корреляционным отношением η_{yx} называется величина, определяемая равенством

$$\boxed{\eta_{yx} = \sqrt{\frac{\left(\sum n_x (\bar{y}_{xi} - \bar{y})^2\right)/n}{\left(\sum n_y (y - \bar{y})^2\right)/n}}} \qquad (11)$$

где n — объем выборки,

n_x — частота значения x признака X,

n_y — частота значения у признака Y,

$\bar{y}$ — общая средняя признака Y,

$\bar{y}_{xi}$ — условная средняя признака Y.

Аналогично определяется выборочное корреляционное отношение η_{xy}

$$\boxed{\eta_{xy} = \sqrt{\frac{\left(\sum n_y (\bar{x}_{yi} - \bar{x})^2\right)/n}{\left(\sum n_x (x - \bar{x})^2\right)/n}}} \qquad (12)$$

ЗНАЧЕНИЕ КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ.

По коэффициенту корреляции r можно судить о наличии и тесноте линейной корреляционной связи между признаками X и Y. По корреляционным отношениям можно судить только о наличии и силе корреляционной связи между признаками X и Y, но не о форме связи, которая устанавливается из геометрических соображений.

Свойства корреляционных отношений

1. Корреляционное отношение заключено между 0 и 1, то есть

$$0 \le \eta_{yx} \le 1 .$$

2. Если корреляционная связь между признаками X и Y отсутствует, то $\eta_{yx} = 0$ и обратно.

3. Если $\eta_{yx} = 1$, то между признаками X и Y существует обычная функциональная связь.

4. Чем ближе значение η_{yx} к 1, тем сильнее корреляционная связь между признаками X и Y, а чем ближе η_{yx} к 0, тем слабее эта зависимость.

5. Если $\eta_{yx} = |r|$ регрессия y на x является линейной.

6. Коэффициент линейной корреляции r не превосходит по модулю $\eta_{yx} = 1$ то есть $|r| \le \eta_{yx}$.

Теснота связи между признаками X и Y при любой форме корреляции может быть измерена с помощью **индекса корреляции** i.

Если опытные данные не сгруппированы в корреляционную таблицу, то индекс корреляции находят по формуле

$$\boxed{i = \sqrt{1 - \frac{S_{yx}^2}{S_y^2}}} \qquad (13)$$

где $\boxed{S_{yx}^2 = \left(\sum (y_j - \bar{y}_x)^2\right) / n}$ — средний квадрат отклонений фактических значений y от значений y, вычисленных по уравнению регрессии;

$\boxed{S_y^2 = \left(\sum (y_j - \bar{y}_x)^2\right) / n}$ — средний квадрат отклонений фактических значений y от их средней арифметической.

Если опытные данные сгруппированы в корреляционную таблицу, то индекс корреляции находят по формуле

$$\boxed{i = \sqrt{1 - \frac{Q_e}{Q}}} \quad (14)$$

$$\boxed{i = \sqrt{\frac{\sigma_y^2}{S_y^2}}} \quad (15)$$

где $Q_e = Q - Q_R$, $Q = \sum (y_i - \bar{y})^2$, $Q_R = \sum (\bar{y}_{xi} - \bar{y})^2$,

$S_y^2 = (\sum (y_i - \bar{y})^2 n_j)/n$, $\sigma_y^2 = \sum (\bar{y}_x - \bar{y})^2 \cdot n_j / n$.

Индекс корреляции по величине изменяется от 0 до 1. По индексу корреляции можно определять, как правило, тесноту связи между признаками X и Y, но не обязательно форму криволинейной связи.

Проверка адекватности модели

Проверить адекватность модели — это значит установить, соответствует ли построенное уравнение регрессии опытным данным и достаточно ли включенных в уравнение факторных признаков X_i для описания результативного признака. Оценка значимости уравнения регрессии производится на основе дисперсионного анализа. В случае однофакторной нелинейной регрессии находят статистику $F_н$ по формуле

$$\boxed{F_{\text{н}} = \frac{Q_R(n-2)}{Q_e}} \qquad (16)$$

Оценке подвергаются не два, а k параметров уравнения регрессии, то находят статистику

$$\boxed{F_{\text{н}} = \frac{Q_R(n-2)}{Q_e(k-1)}} \qquad (17)$$

Затем при заданном уровне значимости α и числах степеней свободы $k_1 = 1$, $k_2 = n-2$ по таблице критических точек распределения Фишера-Снедекора находят $F_т = F_{\alpha; k_1; k_2}$. Если $F_н > F_т$, то модель регрессии согласуется с опытными данными, в противном случае нет.

Лабораторная работа № 5

Цель работы: овладение способами выбора модельного уравнения нелинейной регрессии, выработка умения и навыков расчета параметров уравнения, проверка его надежности.

Содержание работы: На основании опытных данных требуется:

1. Построить корреляционное поле. По характеру расположения точек в корреляционном поле подобрать вид функции регрессии.
2. Написать уравнение функции регрессии.
3. Определить тесноту корреляционной связи между рассматриваемыми признаками.
4. Проверить адекватность модели.
5. Построить кривую регрессии в системе координат.

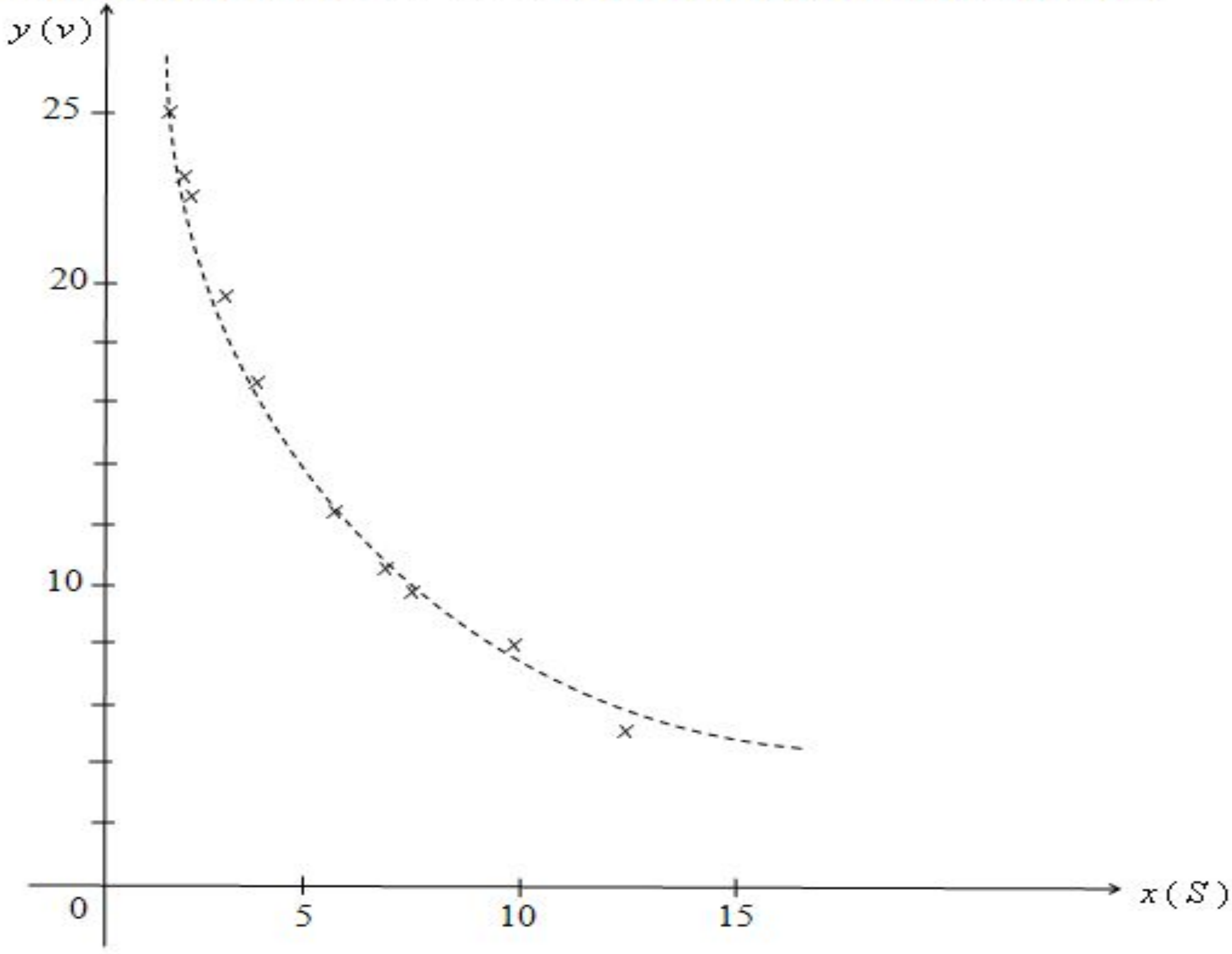
Задача. При обработке металлов резанием устанавливается зависимость резания металла от различных характеристик резца и стружки. Зависимость скорости резания v (м/мин.) и площади поперечного сечения стружки S (мм^2) при обработке хромоникелевой стали задана табл. 3.

Таблица 3

$x = S$	1,1	1,4	1,7	2,1	2,6	4,7	6,1	7,0	10	12,8
$y = v$	25	22,7	22,1	19,8	17	12,3	10,7	10	8,2	6,7

Выполнение работы

В системе координат $0xy$ $(0Sv)$ строим корреляционное поле



По расположению точек в корреляционном поле видно, что около них можно провести ветвь гиперболы. Следовательно, уравнение функции регрессии будем искать в виде (4) или (6). Для выбора одного из этих уравнений применим необходимые условия.

Для формулы $y = a + \frac{b}{x}$ по табл. 1 проверяем выполнение равенства $y(\frac{2x_1 x_n}{x_1 + x_n}) = \frac{y(x_1) + y(x_n)}{2}$.

$$y\left(\frac{2x_1 x_n}{x_1 + x_n}\right) = y\left(\frac{2x_1 x_{10}}{x_1 + x_{10}}\right) = y\left(\frac{2 \cdot 1{,}1 \cdot 12{,}8}{1{,}1 + 12{,}8}\right) = y(2{,}02).$$

Значение $y(2{,}02)$ находим линейным интерполированием по формуле (10):

$$y(2{,}02) = y(1{,}7) + \frac{y(2{,}1) - y(1{,}7)}{2{,}1 - 1{,}7}(2{,}02 - 1{,}7) = 22{,}1 + \frac{19{,}8 - 22{,}1}{0{,}4} \cdot 0{,}32 = 20{,}26.$$

$$\frac{y(x_1) + y(x_n)}{2} = \frac{y(1{,}1) + y(12{,}8)}{2} = \frac{25 + 6{,}7}{2} = 15{,}85.$$

Вычисляем отклонение Δ_1: $\Delta_1 = |15{,}85 - 20{,}26| = 4{,}41$.

Для формулы $y = \frac{1}{ax+b}$ проверяем выполнение равенства $y = \left(\frac{x_1 + x_n}{2}\right) = \frac{2\,y(x_1)\cdot y(x_n)}{y(x_1) + y(x_n)}$.

$$y = \left(\frac{x_1 + x_n}{2}\right) = y\left(\frac{x_1 + x_{10}}{2}\right) = y\left(\frac{1{,}1 + 12{,}8}{2}\right) = y(6{,}95).$$

Значение $y(6{,}95)$ находим линейным интерполированием:

$$y(6{,}95) = y(6{,}1) + \frac{y(7) - y(6{,}1)}{7 - 6{,}1}(6{,}95 - 6{,}1) = 10{,}7 + \frac{10 - 10{,}7}{0{,}9}\cdot 0{,}85 = 10{,}04.$$

$$\frac{2\,y(x_1)\cdot y(x_n)}{y(x_1) + y(x_n)} = \frac{2\,y(1{,}1)\cdot y(12{,}8)}{y(1{,}1) + y(12{,}8)} = \frac{2\cdot 25\cdot 6{,}7}{25 + 6{,}7} = 10{,}57$$

Вычисляем отклонение Δ_2: $\Delta_2 = |10{,}57 - 10{,}04| = 0{,}53$. Так как $\Delta_2 = 0{,}53 < \Delta_1$, то по методу необходимых условий выбираем формулу $y = \frac{1}{ax+b}$.

Произведем выбор одной из выше рассматриваемых формул по методу конечных разностей. Пусть $\boxed{y = a + \frac{b}{x}}$. Сводим эту зависимость к линейной $Y = AX + B$, где $X = x$, $Y = xy$ (смотри таблицу). Вычисляем отношения $\Delta Y / \Delta X$. Составляем расчетную таблицу 4.

Таблица 4

$X = x$	1,1	1,4	1,7	2,1	2,6	4,7	6,1	7,0	10	12,8
y	25	22,7	22,1	19,8	17	12,3	10,7	10	8,2	6,7
$Y = xy$	27,5	31,78	37,57	41,58	44,2	57,81	65,27	70	82	85,76
ΔY	4,28	5,79	4,01	2,62	13,61	7,46	4,73	12	3,76	
ΔX	0,3	0,3	0,4	0,5	2,1	1,4	0,9	3	2,8	
$\frac{\Delta Y}{\Delta X}$	14,27	19,3	10,02	5,24	6,48	5,33	5,26	4	1,34	

Рассмотрим зависимость $\boxed{y = \frac{1}{ax+b}}$. Пользуясь таблицей 1, сводим нелинейную зависимость к линейной $Y = AX + B$, где $X = x$, $Y = \frac{1}{y}$. Для нахождения отношений $\Delta Y / \Delta X$ составляем расчетную таблицу 5.

Таблица 5

$X = x$	1,1	1,4	1,7	2,1	2,6	4,7	6,1	7,0	10	12,8
y	25	22,7	22,1	19,8	17	12,3	10,7	10	8,2	6,7
$Y = \frac{1}{y}$	0,04	0,044	0,045	0,05	0,059	0,081	0,093	0,1	0,12	0,15

ΔY	0,004	0,001	0,005	0,009	0,022	0,012	0,007	0,02	0,03
ΔX	0,3	0,3	0,4	0,5	2,1	1,4	0,9	3	2,8
$\Delta Y / \Delta X$	0,013	0,003	0,012	0,018	0,01	0,008	0,008	0,007	0,011

ВЫВОД. Отношения $\frac{\Delta Y}{\Delta X}$, полученные для формулы $y = \frac{1}{ax+b}$ мало отличаются друг от друга, чем для формулы $y = a + \frac{b}{x}$. Поэтому по методу конечных разностей в качестве лучшей выбираем формулу $y = \frac{1}{ax+b}$. К такому же выводу мы пришли, применяя метод необходимых условий. Итак, зависимость скорости резания от площади поперечного сечения стружки при обработке хромоникелевой стали выражается формулой $\hat{y}_x = \frac{1}{ax+b}$.

Оценки a и b неизвестных параметров истинного уравнения регрессии находим, решая систему нормальных уравнений (7):

$$\begin{cases} a_0 n + a_1[x] = \left[\frac{1}{y}\right] \\ a_0[x] + a_1\left[x^2\right] = \left[\frac{x}{y}\right] \end{cases}$$

Для начисления сумм, входящих в систему, составляем расчетную табл. 6.

Таблица 6

x	y	$1/y$	x/y	x^2
1,1	25	0,04	0,044	1,21
1,4	22,7	0,044053	0,061674	1,96
1,7	22,1	0,045549	0,076923	2,89
2,1	19,8	0,050505	0,106061	4,41
2,6	17	0,058824	0,152941	6,76
4,7	12,3	0,081301	0,382114	22,09
6,1	10,7	0,093458	0,570093	37,21
7,0	10	0,1	0,7	49
10	8,2	0,121951	1,21951	100
12,8	6,7	0,149254	1,910448	163,84
39,6		0,784595	5,223764	389,37

Решаем систему уравнений:

$$\begin{cases} 10a_0 + 39{,}6a_1 = 0{,}784595 \\ 39{,}6a_0 + 389{,}37a_1 = 5{,}223764 \end{cases}, \quad a_0 = 0{,}04282, \quad a_1 = 0{,}009.$$

Уравнение регрессии примет вид:

$$\hat{y}_x = \frac{1}{0{,}09x + 0{,}04282}.$$

Оценим силу корреляционной связи между скоростью резания и площадью поперечного сечения стружки хромоникелевой стали. Вычислим индекс корреляции по формуле (13):

$$i = \sqrt{1 - \frac{\hat{S}^2_{yx}}{\hat{S}^2_y}}, \quad \text{где} \quad \hat{S}^2_{yx} = \frac{1}{n-1}\sum(y_i - \bar{y}_x)^2, \quad \hat{S}^2_y = \frac{1}{n-1}\sum(y_i - \bar{y})^2$$

(так как $n = 10 < 50$).

Для нахождения S^2_{yx} и $\hat{S}^2_y$ составляем расчетную табл. 7.

Таблица 7

x_i	y_i	$\hat{y}_x$	$(y-\hat{y}_x)^2$	$(y-\bar{y})^2$
1,1	25	19	36	91,2025
1,4	22,7	18	22,09	52,5625
1,7	22,1	17,2	24,01	44,2225
2,1	19,8	16,2	12,96	18,9225
2,6	17	15,1	3,61	2,4025
4,7	12,3	11,7	0,36	9,9225
6,1	10,7	10,2	0,25	22,5625
7,0	10	9,4	0,36	29,7025
10	8,2	7,5	0,49	52,5625
12,8	6,7	6,3	0,16	76,5625
			100,29	400,625

Тогда $i = \sqrt{1 - \frac{11,14333}{44,45139}} = 0,86$

Связь между скоростью резания и площадью поперечного сечения стружки хромоникелевой стали сильная.

Проверяем адекватность полученного уравнения регрессии по **критерию Фишера-Снедекора**. Находим статистику

$$F_{\text{н}} = \frac{i^2(n-2)}{1-i^2} = \frac{0{,}86^2(10-2)}{1-0{,}86^2} = 22{,}7\ .$$

При уровне значимости $\alpha = 0{,}05$ и числах степеней свободы $k_1 = \mathbf{1}$, $k_2 = n - 2 = 10 - 2 = 8$ по таблице критических точек распределения Фишера-Снедекора находим $F_{\text{т}} = F_{\alpha;k_1;k_2} = F_{0,05;1;8} = 5{,}32$. Так как $F_{\text{н}} = 22{,}7 > 5{,}32$, то модель адекватна. Следовательно, зависимость скорости резания от площади поперечного сечения стружки при обработке хромоникелевой стали по данным выборки описывается уравнением $\hat{y}_x = \dfrac{1}{0{,}009x + 0{,}04282}$.